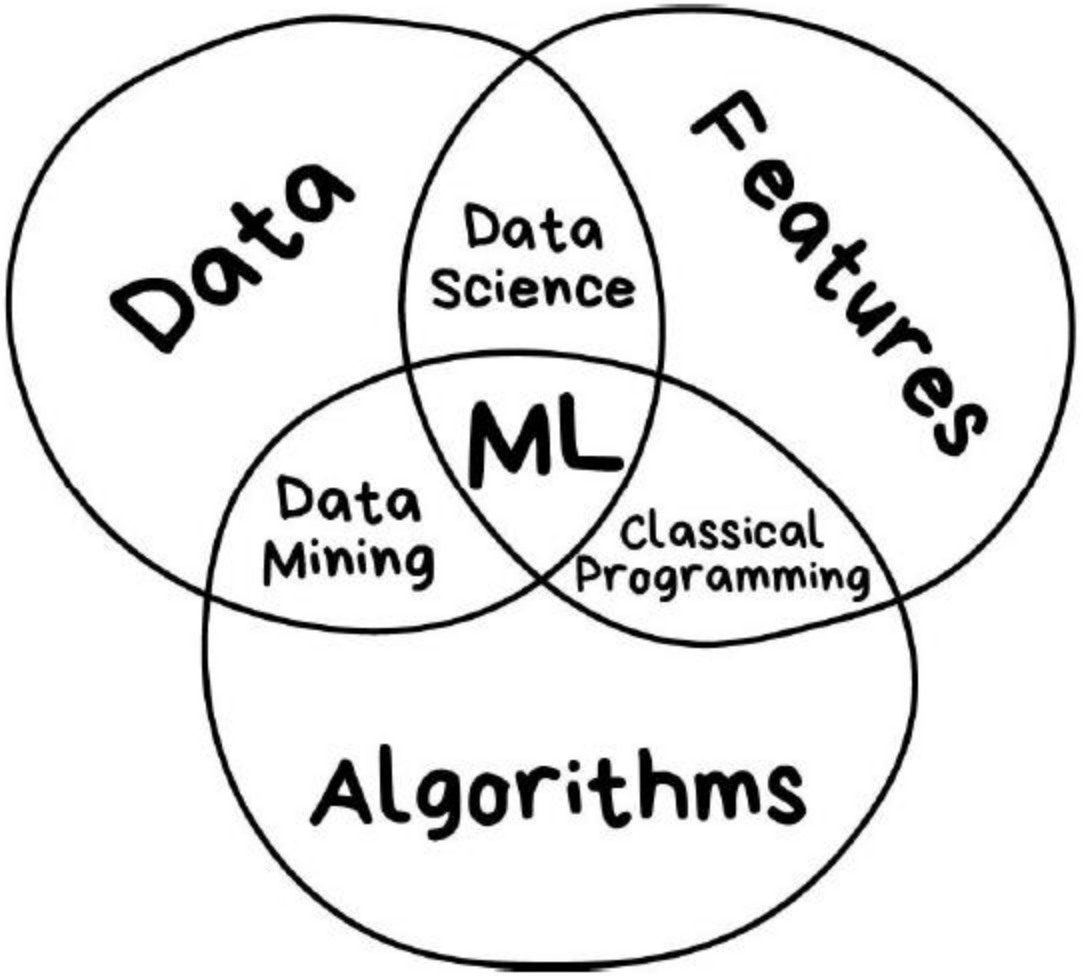
# Predictive Analytics

Abhimanyu Gupta

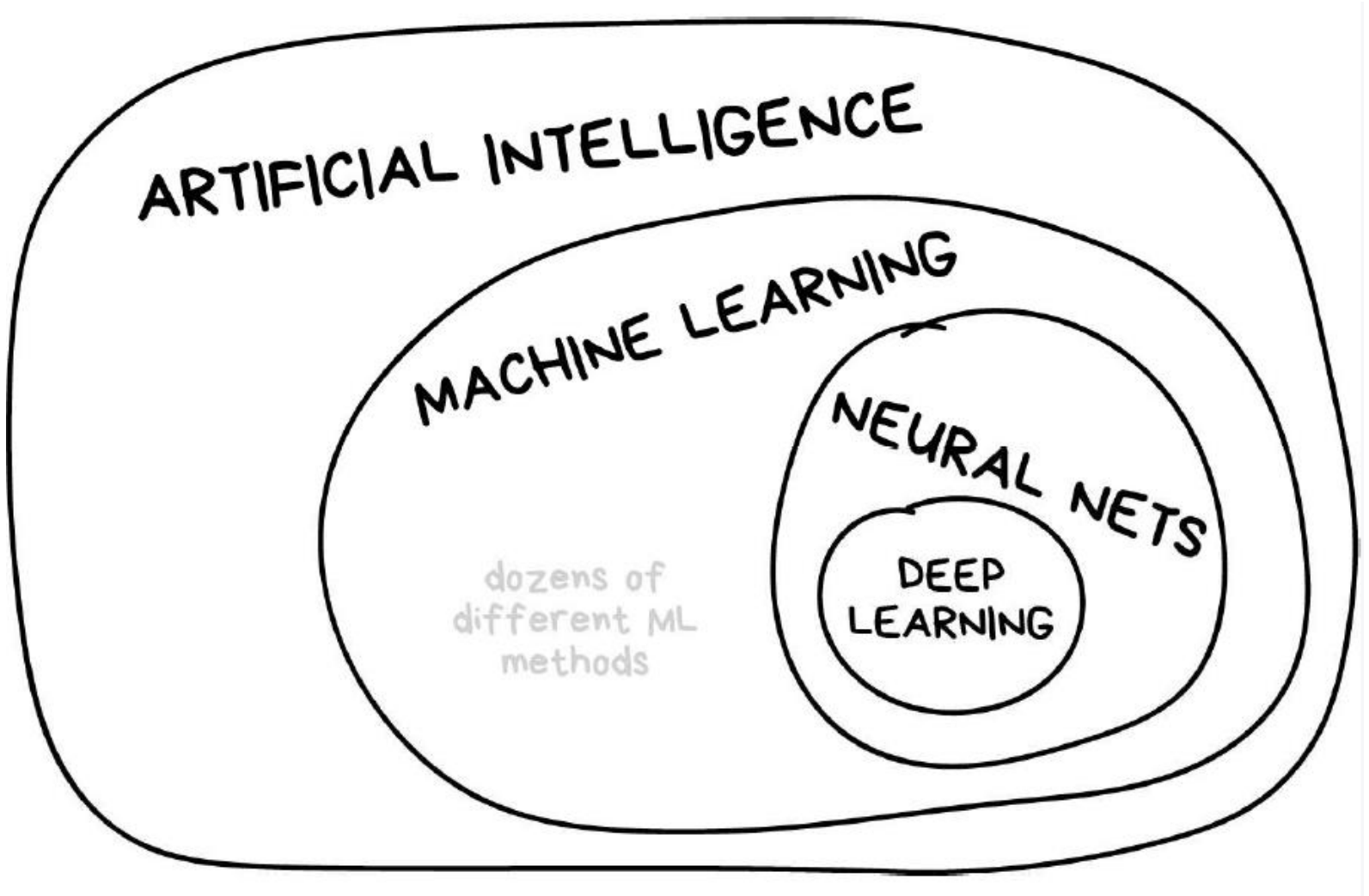
# What is Predictive Analytics?

# Applications of Predictive Analytics

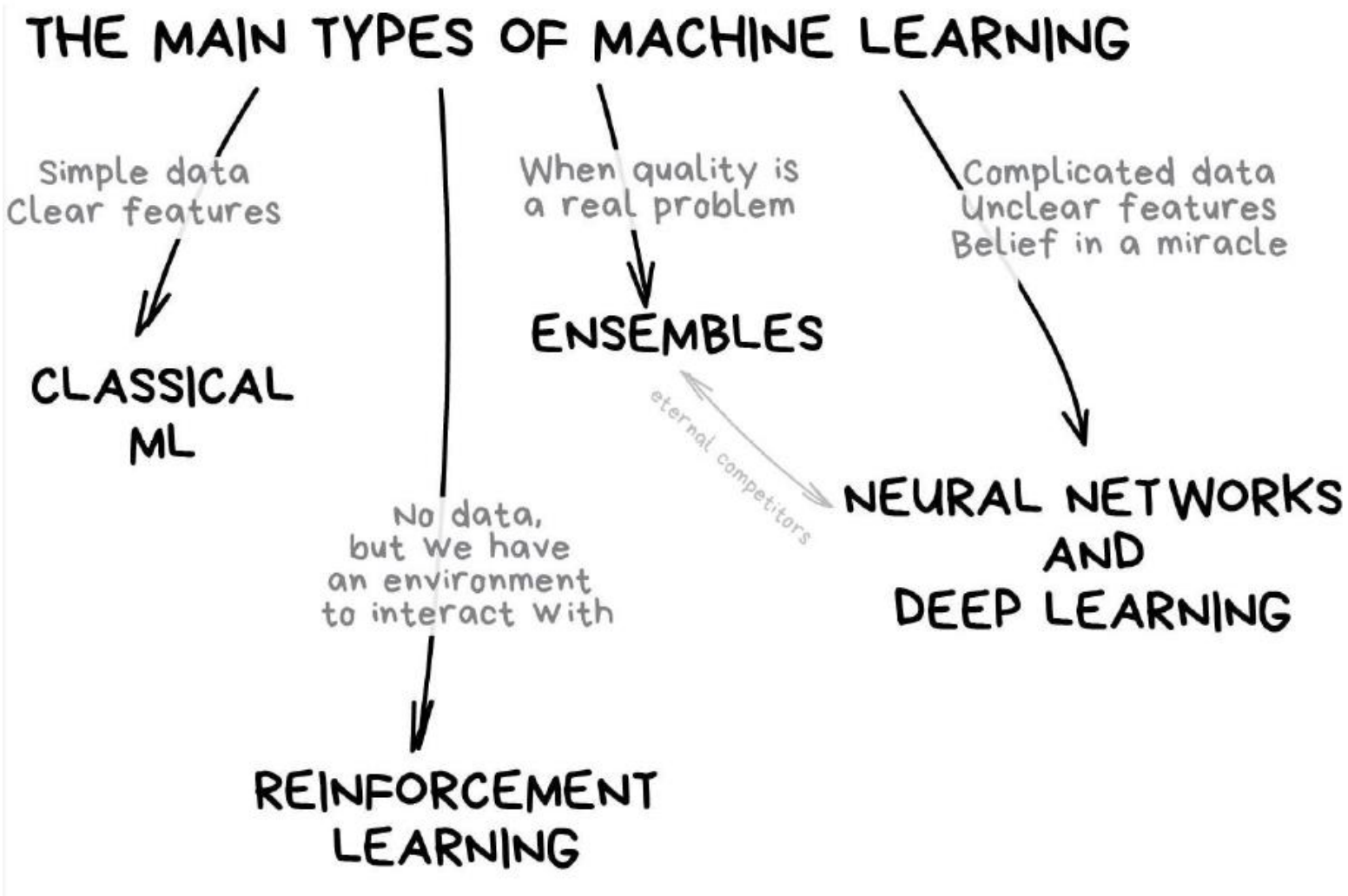
# What is Machine Learning?



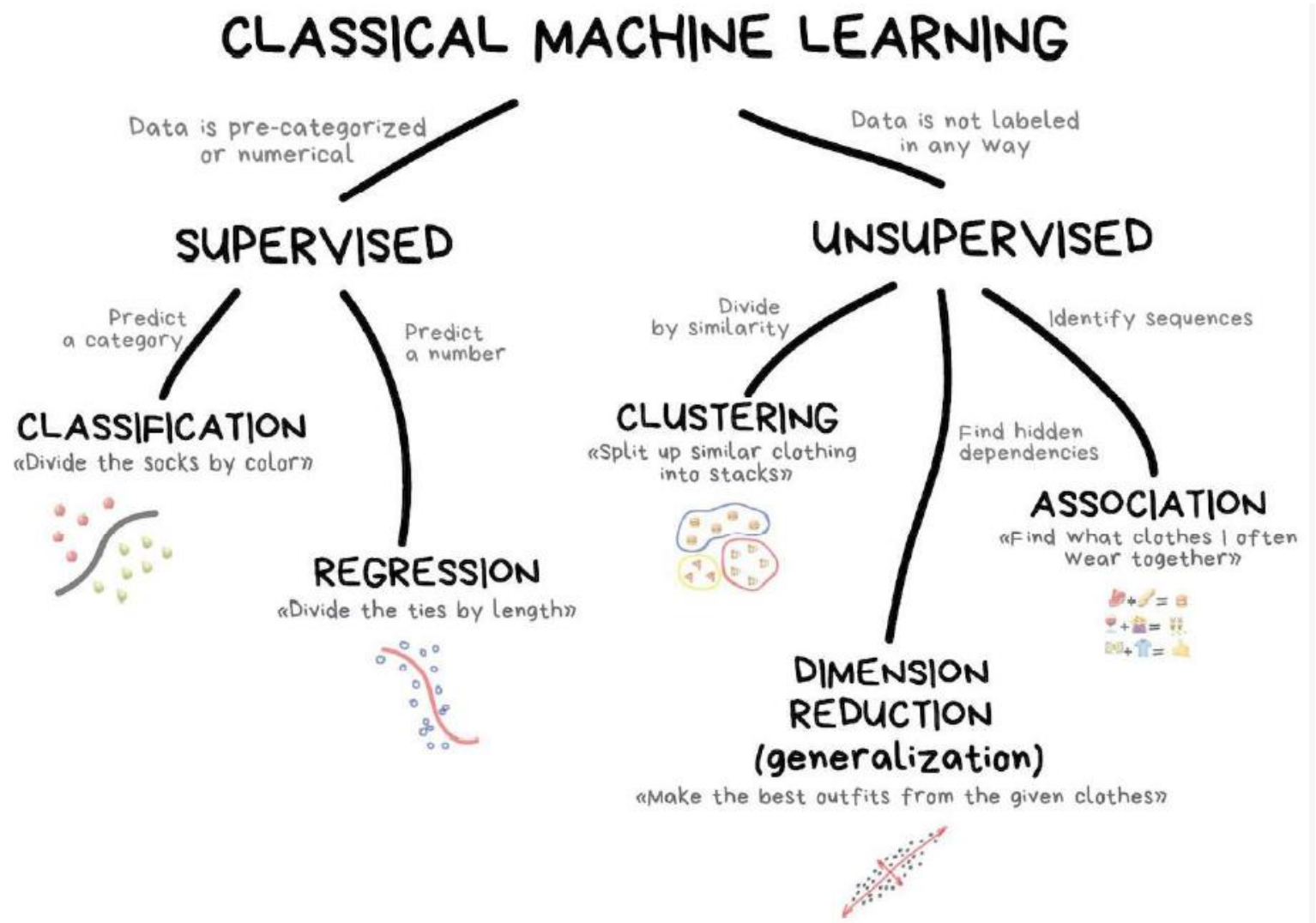
# What is Machine Learning?



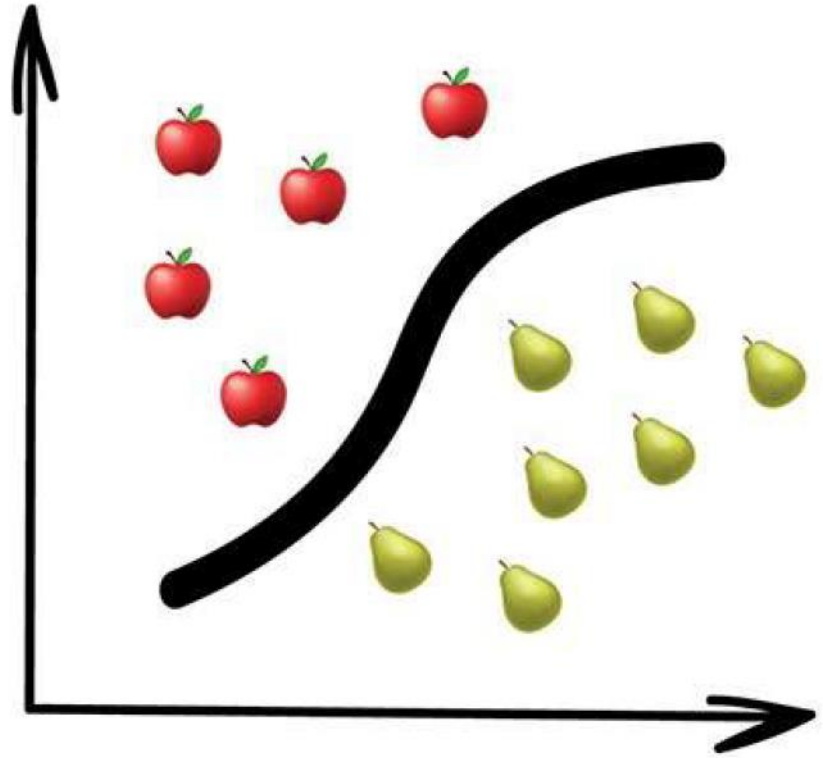
# Types of Machine Learning

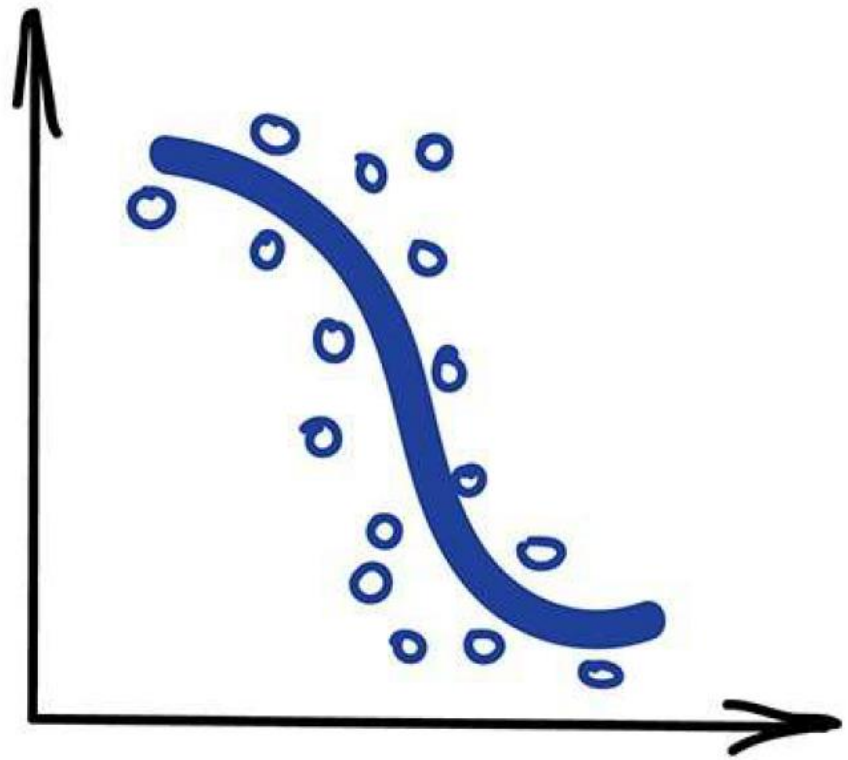


# Classical Machine Learning



# Supervised Learning

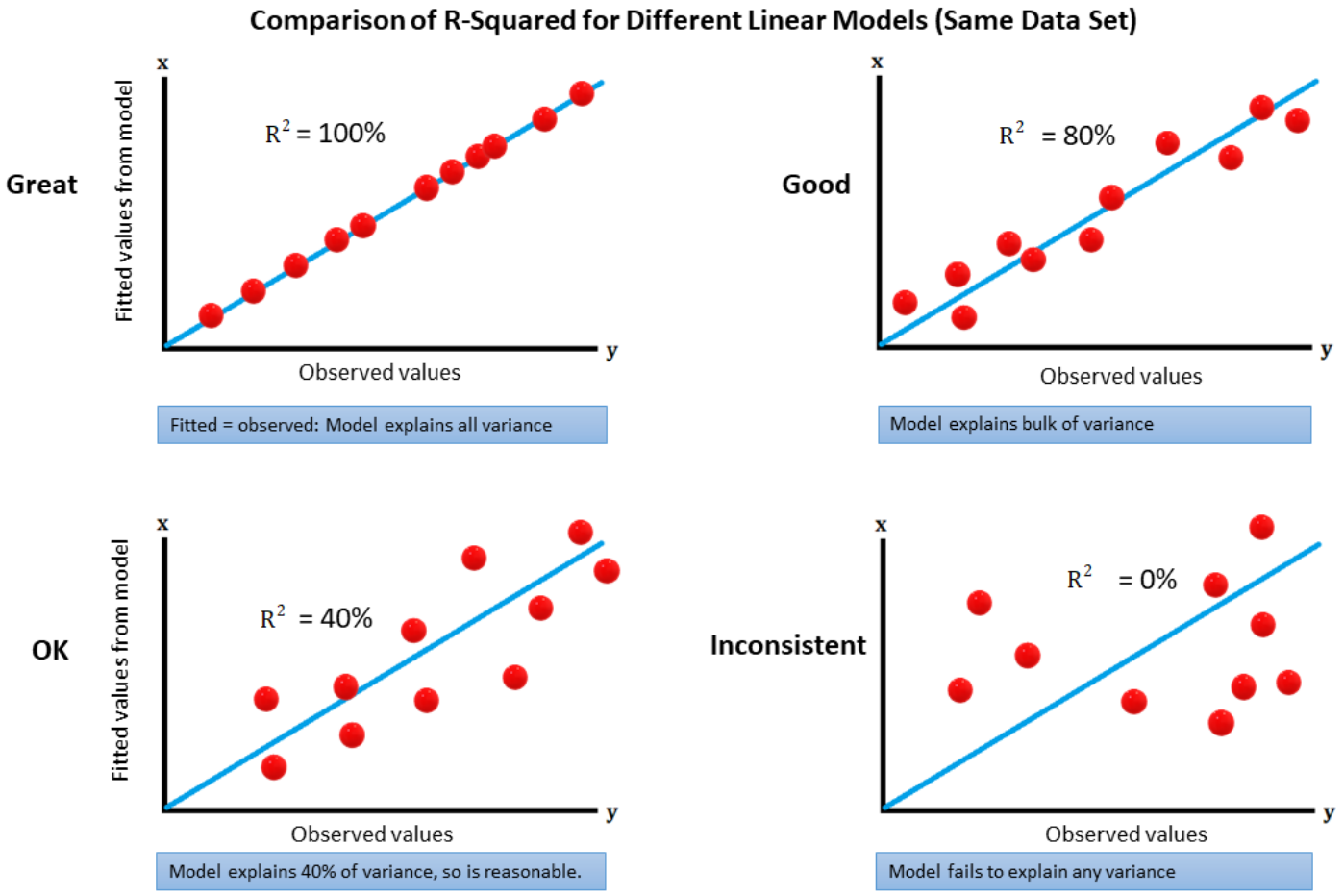


# Supervised Learning Algorithms

| Algorithm | Problem Type | Assumptions | Results interpretable? | Easy to explain? | Predictive accuracy | Training speed | Prediction speed | Amount of parameter tuning? | Performance with small datasets? |
|---|---|---|---|---|---|---|---|---|---|
| **Linear Regression** | Regression | Normality & Linearity | Yes | Yes | Lower | Fast | Fast | None | Yes |
| **Logistic Regression** | Classification | Binary classes only! | Somewhat | Somewhat | Lower | Fast | Fast | None | Yes |
| **Linear Discriminant Analysis** | Classification | Normality | Yes | Yes | Lower | Fast | Fast | None | Yes |
| **Decision trees** | Either | None | Somewhat | Somewhat | Lower | Fast | Fast | Some | No |
| **Random Forests** | Either | None | A little | No | Higher | Slow | Moderate | Some | No |
| **k-Nearest Neighbor** | Classification | None | Yes | Yes | Lower | Fast | Depends on k | Minimal | No |
| **Support Vector Machine** | Either | None | Yes | Yes | Lower | Fast | Fast | Some | No |
| **Naive Bayes** | Classification | None | Somewhat | Somewhat | Lower | Fast | Fast | Some | Yes |
| **Neural networks** | Either | None | No | No | Higher | Slow | Fast | Lots | No |

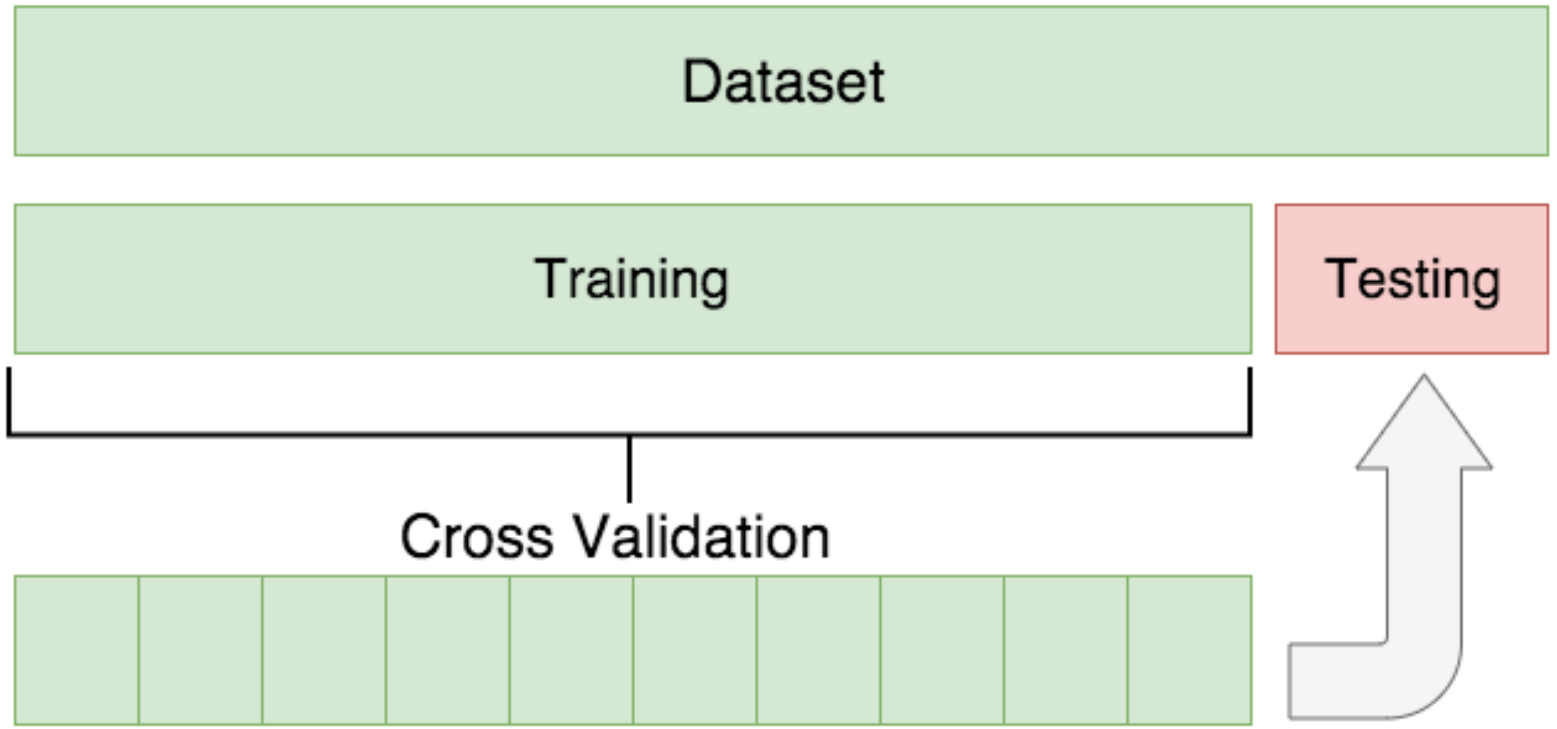
# Regression Accuracy with R-squared



# Classification Accuracy with Confusion Matrix

| n=165 | **Predicted: NO** | **Predicted: YES** | |
|---|---|---|---|
| **Actual: NO** | TN = 50 | FP = 10 | 60 |
| **Actual: YES** | FN = 5 | TP = 100 | 105 |
| | 55 | 110 | |

# Training & Testing Data



# Steps for Prediction

```
Choose a data set
Load the data set
Select the dependent variable
Select all independent variables
Select a subset of the data set as training data
Determine testing data as the complementary subset of the training data
If the dependent variable is numeric then
          Fit a Regression model on the training data
          Predict the value of dependent variable in the training data using the fitted model
          Determine accuracy as the R-squared value between actual and predicted values of dependent variable
If the dependent variable is NOT numeric then
          Fit a Classification model on the training data
          Predict the class of dependent variable in the training data using the fitted model
          Create a confusion matrix to determine the actual and predicted classes of the training data
          Determine accuracy as the proportion of correct predictions of dependent variable
Predict the value of dependent variable in the testing data using the fitted model
Save the testing data, with predicted values, in the local drive
```